Graduate Research in Engineering and Technology (GRET)

Volume 1 | Issue 2 Article 12

October 2013

AUTOMATED SENTIMENT ANALYSIS AN AUTOMATED ANALYSIS OF NEWS FEEDS

KARTHIK BALASUBRAMANIAN
Department of Information Technology, SRM University, Kattangulathur, Chennai - 603203,
bks4line@gmail.com

APARAJITH CHANDRAN
Department of Information Technology, SRM University, Kattangulathur, Chennai - 603203,
aparajith.chandran@gmail.com

AUTOMATED SENTIMENT ANALYSIS AN AUTOMATED ANALYSIS OF NEWS FEEDS

KARTHIK BALASUBRAMANIAN[1], APARAJITH CHANDRAN[2]

[1,2]Department of Information Technology, SRM University, Kattangulathur, Chennai - 603203
E-mail: bks4line@gmail.com, aparajith.chandran@gmail.com

Abstract- This Paper explains the importance of Sentiment Analysis in today's business. Information which is hidden as unstructured data in the Internet can be utilized more efficiently. In this paper we quote an approach which explains the experiment for collection of news data and analyzing the sentiments for those data. The results provide almost accurate analysis outcomes with a few discrepancies. These are also explained and research is in progress towards making an efficient system.

Keywords- *Sentiment Analysis, Document Summarizer, Data Crawler*

I. INTRODUCTION

Newspapers have become an essential part of modern life. We start our day's work with finding out what's actually happening in the world. We get varieties of information from newspapers ranging from Business and politics to sports and arts. Nowadays each and every newspaper industry has its own news repository on internet like archives. These archives can be utilized more efficiently so that it may improve our quality of life. In a country like India, where the number of internet users is on a rise, this could be put to good use. For example, during an election consistent analysis of people's comments online might provide a better survey of a political leader among the people.

This paper explains the how this kind of analysis is done from the semi structured informative archives from the news repositories and how efficient it is when compared to the manual feedback. Section II reviews the existing research in Sentiment Analysis. Section III reviews the Experiment Procedures. Section IV reviews the Challenges faced during the course of the project. Section V reviews the Results obtained and future Developments. Section VI presents the Conclusion and Acknowledgement.

II. EXISTING SENTIMENT ANALYSIS RESEARCH

A. Analysis on Social Media Feeds and Comments

The whole idea of Sentiment Analysis was derived from the power of social media which can produce great results. Some of the beneficial results that can be concluded by monitoring Social media are:

- Voice of the Voter: Sentiment Analysis will help political organizations, campaigns and news analysts to better understand the issues and positions that matter most to voters. The technology was applied during the 2012 U.S. presidential campaigns which produced accurate results.

Brand Reputation Management is another application of monitoring (text analytics, presentation) technologies to online and social media. These are places where any past, current or prospective customer can post opinions that can damage or boost a brand's reputation.

B. Opinion Analysis:

This is another existing field in Sentiment Analysis. These days, users' opinions and ratings drive the search engine to produce relevant results. Google Analytics provides search solutions based on the opinions and priorities of the user.

For example, if you are looking to buy a DSLR camera online, you might find in a few moments advertisements regarding the camera on your web browser. This is also called Opinion Mining.

III. EXPERIMENT DESCRIPTION

Sentiment analysis can be done in the above ways as stated in Section III. But problems faced because of the inaccuracy renders the experiments useless. For example, if a user discredits a particular product based on personal experiences on social networking sites and blogs, it would spoil the reputation of that product and likewise the company pertaining to his/her comment. So to get legitimate data about a product, person or business, it is important to cite a proper source.

That is the reason this experiment deals with finding the sentiments from legitimate source of news feeds where one can get unbiased information. Although Sentiment Analysis does indeed register the positive and negative sentiments, it cannot identify sarcastic statements (that is to say these get classified as positive or negative too).

A. Steps Involved in Automated Sentiment Analysis

Automated Sentiment Analysis involves the following steps.

- Data Collection

The data from the news sites must be collected and its relevance ensured. So we need to crawl and populate the database with different column names, segregating date, tags, title and content. Any open source database can be used. It can be an SQL database like MYSql or a Nosql Database like mongoDb. So, once we collect the data it needs to be arranged in an orderly manner according to date and time. Care must be taken to crawl only the relevant links and not advertisements.

- Querying the Database

Now that the data is collected, the users are given the freedom to choose the content they want to search and analyze the sentiments. Once the user gives the input, all the data which contain the particular search input are collected. Before the search process is initiated, the search string is classified into different phrases. If alternative meanings of those phrases are found in the database, those data are also collected.

- Summarization of Data

The data collected prives too large to be analyzed for sentiments. So when the user provides a search string, not only the data that contains the particular phrase, but also the synonyms of that particular phrase are collected. Those huge data have to be normalized or summarized to some important data that the phrases emphasize on. The summarized data have to undergo sentiment analysis. During the summarization process, we ensured that important data tags are never missed out and unnecessary data are eliminated from the database.

- Sentiment Analysis

Now we are left with the summarized data which might be not more than 300 words. Using this, we can group the search results into positive, negative and neutral news contents. This would be mostly valid for proper noun searches e.g. personalities or products. But may not provide efficient search results for verbs, adjectives etc.

Ways to perform Sentiment Analysis

There are 2 approaches to the Sentiment Analysis. They are

1. General Inquirer Approach
2. Rule Based Classifiers

B.1.General Inquirer based Approach

After we obtain the summarized data, its content tags are weighed against each and every word in a dictionary. This dictionary contains a pool of positive, negative and neutral words. So by analyzing each and every tag of the summarized document, we might detect the percentage of positivity or negativity by counting the number of positive or negative tags. But the assurance of perfect analysis cannot be given for this type of experiment because the document might emphasize positive statements but the result might be negative. So there are inconsistencies in this model. Another disadvantage is that the General Inquirer Method might not reveal the important words for sentiment classification.

B.2.Rule Based Classifiers

In this approach too, there exists a dictionary that contains all the positive and negative words. But the implementation varies. Each and every sentiment result is based on a set of structured loop of rules which is pre-programmed and the results derived. It is more resourceful than the previous algorithm. Thus, higher accuracy is obtained using this method. The major advantage of this type of analysis is that if the given object (word) based analysis does not satisfy the rules the analysis is dropped. So coverage of analysis is aptly governed and high accuracy rates are achieved.

There are various other methods to carry out sentiment analysis. There are also hybrid analysis models that further enhance the efficacy of sentiment analysis.

IV. CHALLENGES FACED DURING IMPLEMENTATION

- During data collection, unnecessary links were being crawled.
- We encountered performance issues while crawling. Sometimes irrelevant links were being crawled. This lead to errors.
- User Based Search also had problems. For example, when the search term “modi” was given, the results yielded all the tags that not only included “modi” but also keywords like “modified”.
- Data Summarization implementation was one of the toughest phases as it required the best summarization methods.
- Sentiment Analysis also had problems. The accuracy of the Sentiment Analysis was not up to the mark initially. Subsequently, many rule based classification changes were made to obtain the desired accuracy.

These are some of the mistakes and errors encountered while doing the project. Each step of implementation is separated as module and the best methods were planned and implemented to rectify these errors.

V. FUTURE DEVELOPMENTS

A GUI based search is being planned. A good user experience will definitely bolster this effort.

1. The experiment is intended as a mobile application for all platforms (including Android, Windows and Apple).
2. The experiment is expected to work for a specific application rather than a general Sentiment Analysis. For example many new fields can be added like Food, Clothing, Tourism etc; data from expert reviews in news and signature sites for those particular fields collected and the sentiments about those particular fields analyzed. For example if a user searches for a hotel using this application, the reviews on that hotel will be assembled and sentiments of those reviews from experts can be analyzed leading to a more trimmed and tailored search outcome.
3. Food and Health Analytics are two booming fields that have no risks when the sentiments are analyzed. Rather, they are used for upgradation of environment. Therefore these are the two fields the focus would be on in the later stages of the project.

VI. RESULTS OBTAINED

From the above research, each and every word's strength is analyzed using a framework that provides sentiment called SentiWordNet an extensive product of WordNet dictionary from Princeton University of America. We used such a product because the predecessor WordNet is one of the signature product which brought about a revolution in the field of linguistic analysis. With the SentiWordnet the accuracy of analysis is more and the efficiency can be increased by the Document Classifier algorithm which would be our next step in development of this project. We obtained 68% of accuracy from the analysis of news feeds.

ACKNOWLEDGMENT

The Project was supervised by our project guide Nimala.K, Professor Data Mining , SRM University. We would like to acknowledge the assistance of Mr. Sundarrajan Senior Product Engineer, Alpha Cloud labs, Chennai, who has been a constant support during the course of development of this project. We would also like to thank Mr. Mahesh Arumugam for providing us the opportunity to work on this project.

CONCLUSION

Due to immense growth in available information on the Web (Internet) the customer is so overwhelmed by the abundant information that he is unable to make a decision from the mixed positive, negative, good, bad and fake reviews/comments by other Internet users. So an unbiased platform of information is needed to do the analysis to make a customer convince on making a decision. That is the reason that motivated us to perform the Sentiment Analysis of News feeds. Our evidence and results are limited to the domain of news feeds, but the ideas in this paper have tremendous potential for future research.

REFERENCE

[1]. Cai, Keke, Spangler, Scott:leveraging Sentiment Analysis for Topic Detection , Publication Year: 2008, Page(s): 265 – 271

[2]. Deneke , Kerstin, Are SentiWordNet scores suited for multi-domain sentiment classification? Fourth International Conference on Digital Information Management, 2009. ICDIM 2009. Publication Year: 2009, Page(s): 1 - 6

[3]. Ghorpade, Tushar, Latha L, Ragha, Featured based sentiment classification for hotel reviews using NLP and Bayesian classification, International Conference on Communication Communication, Information & Computing Technology (ICCICT), 2012.Publication Year: 2012, Page(s): 1-5.

[4]. Neviarouskaya, Alina, Predinger, Helmut, Ishizuka, Mitsuru, IEEE Transactions on Affective Computing, Volume 2, issue-1. Publication year: 2011, Page(s): 22-36.

◈◈◈